Code Generation: Results Achieved with Zero-shot Prompting When Asking for a 5-level Judgment

Table 1: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment.

LLM	Code generation			
	Java		Python	
	#	%	#	%
DeepSeek Coder 1.3B	38	2.70%	13	1.01%
DeepSeek Coder 6.7B	0	0.00%	7	0.55%
DeepSeek Coder 33B	2	0.14%	45	3.51%
CodeLlama 7B	1	0.07%	103	8.04%
CodeLlama 13B	0	0.00%	460	35.91%
CodeLlama 34B	1	0.07%	52	4.06%
GPT-3.5-turbo	0	0.00%	0	0.00%
GPT-4-turbo	0	0.00%	0	0.00%

	DSC 1.3B	**DSC 6.7B**	**DSC 33B**	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
Java	0.05	-0.07	0.07	-0.16	0.06	0.22	0.12	0.25
Python	0.02	0.01	0.05	-0.06	-0.01	0.03	0.07	0.12

Table 2: Code Generation: Spearman correlation coefficient between the 5-level judgments of the LLMs and the pass/fail (0 or 1) ground truth. The Kappa agreement is not shown because the judgments of the LLMs range from 1 to 5, whereas the ground truth is binary.

	DSC 1.3B	**DSC 6.7B**	**DSC 33B**	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**	**Human Written**	**Own *vs* LLMs**	**Own *vs* LLMs \ F**	**Own *vs* Human**
DSC 1.3B	3.97	4.00	4.00	4.00	3.97	4.00	3.98	4.00	3.91	(N)	(N)	(N)
DSC 6.7B	5.00	4.97	4.88	4.97	4.97	4.96	5.00	5.00	4.69	(N)	(N)	** (S)
DSC 33B	4.78	4.84	4.84	4.73	4.89	4.88	4.89	4.92	4.77	(N)	(N)	(N)
CL 7B	4.47	4.28	4.04	4.42	4.37	4.54	4.35	4.28	4.06	(N)	(S)	*** (M)
CL 13B	3.97	3.97	3.96	3.97	3.95	3.94	4.00	3.98	4.01	(N)	(N)	(N)
CL 34B	3.78	4.12	3.76	4.12	3.95	3.46	4.18	4.12	3.36	** (S)	** (S)	(N)
GPT-3.5	4.81	4.91	4.64	4.73	4.77	4.73	4.89	4.85	4.06	(N)	(N)	*** (M)
GPT-4	4.41	4.72	4.56	4.12	4.71	4.38	4.93	4.97	3.59	** (S)	*** (S)	*** (L)
Average (all)	**4.40**	**4.48**	**4.33**	**4.38**	**4.45**	**4.36**	**4.53**	**4.51**	**4.05**	-	-	-
Average (large)	**4.35**	**4.51**	**4.35**	**4.33**	**4.46**	**4.28**	**4.58**	**4.57**	**3.95**	-	-	-

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 3: **(Java)** 5-level scenario: average rating given by the judge LLM (rows) to the functions generated by the generator LLM or manually written by humans (columns). Note that here only functions **passing** the tests are considered. Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Own *vs* LLMs	Own *vs* LLMs \ F
DSC 1.3B	3.92	3.98	3.95	3.95	3.94	3.93	3.95	3.91	(N)	(N)
DSC 6.7B	4.94	4.96	4.69	4.95	4.95	4.94	4.95	4.95	(N)	(N)
DSC 33B	4.61	4.69	4.76	4.70	4.78	4.79	4.78	4.83	(N)	(N)
CL 7B	4.52	4.47	4.09	4.45	4.37	4.47	4.41	4.35	(N)	(N)
CL 13B	3.92	3.95	3.90	3.95	3.96	3.92	3.98	3.97	(N)	(N)
CL 34B	2.97	3.61	2.98	3.61	3.12	3.16	3.34	3.21	(N)	(N)
GPT-3.5	3.86	4.36	3.96	4.23	4.32	4.27	4.70	4.56	*** (S)	*** (S)
GPT-4	3.12	3.54	3.65	3.51	3.61	3.35	4.22	4.35	*** (S)	*** (M)
Average (all)	**3.98**	**4.19**	**4.00**	**4.17**	**4.13**	**4.10**	**4.29**	**4.27**	-	-
Average (large)	**3.70**	**4.03**	**3.85**	**4.00**	**3.96**	**3.90**	**4.20**	**4.18**	-	-

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 4: **(Java)** 5-level scenario: average rating given by the judge LLM (rows) to the functions generated by the generator LLM or manually written by humans (columns). Note that here only functions **failing** the tests are considered. The *Human Written* column is not present because there are no human written function which fail the tests. Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
DSC 1.3B	4.14	4.00	3.91	4.11	4.09	3.93	4.14	4.06	3.93	(N)	(N)	(S)
DSC 6.7B	4.90	4.90	5.00	5.00	5.00	4.90	4.94	5.00	4.93	(N)	(N)	(N)
DSC 33B	4.75	4.80	4.14	4.59	4.77	4.76	4.46	4.76	4.26	(S)	(S)	(N)
CL 7B	3.89	4.39	4.27	3.87	4.14	4.29	4.10	4.03	4.01	(N)	(N)	(N)
CL 13B	3.83	3.91	4.38	4.00	4.16	4.29	4.33	4.38	4.10	(N)	(N)	(N)
CL 34B	3.81	3.90	4.32	3.78	4.30	4.48	4.29	4.44	4.20	(S)	(N)	(S)
GPT-3.5	3.71	4.43	4.13	4.28	4.41	4.31	4.57	4.34	3.54	(N)	(N)	*** (M)
GPT-4	3.24	4.33	3.65	3.94	4.30	4.03	4.46	4.34	3.13	(N)	(N)	*** (L)
Average (all)	**4.04**	**4.33**	**4.22**	**4.20**	**4.40**	**4.37**	**4.41**	**4.42**	**4.01**	-	-	-
Average (large)	**3.87**	**4.28**	**4.12**	**4.12**	**4.39**	**4.38**	**4.42**	**4.45**	**3.84**	-	-	-

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 5: **(Python)** 5-level scenario: average rating given by the judge LLM (rows) to the functions generated by the generator LLM or manually written by humans (columns). Note that here only functions **passing** the tests are considered. Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Own *vs* LLMs	Own *vs* LLMs \ F
DSC 1.3B	3.99	3.97	4.01	3.92	3.99	3.91	4.02	3.98	(N)	(N)
DSC 6.7B	4.91	4.96	4.93	4.96	4.95	4.92	4.94	4.95	(N)	(N)
DSC 33B	3.94	4.58	4.37	4.21	4.33	4.37	4.41	4.55	(N)	(N)
CL 7B	4.15	4.17	4.25	4.28	4.22	4.19	4.04	4.11	(N)	(N)
CL 13B	4.15	4.07	4.00	4.26	4.05	4.03	4.13	4.31	(N)	(N)
CL 34B	3.71	4.07	4.13	4.13	4.32	4.10	4.26	4.39	(N)	(N)
GPT-3.5	3.28	3.73	3.69	3.46	3.77	3.62	4.01	4.35	(N)	* (S)
GPT-4	2.78	3.37	3.47	2.93	3.27	3.22	3.67	4.28	*** (M)	*** (L)
Average (all)	**3.87**	**4.11**	**4.11**	**4.02**	**4.11**	**4.04**	**4.19**	**4.36**	-	-
Average (large)	**3.57**	**3.96**	**3.93**	**3.80**	**3.95**	**3.87**	**4.10**	**4.38**	-	-

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 6: **(Python)** 5-level scenario: average rating given by the judge LLM (rows) to the functions generated by the generator LLM or manually written by humans (columns). Note that here only functions **failing** the tests are considered. The *Human Written* column is not present because there are no human written function which fail the tests. Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

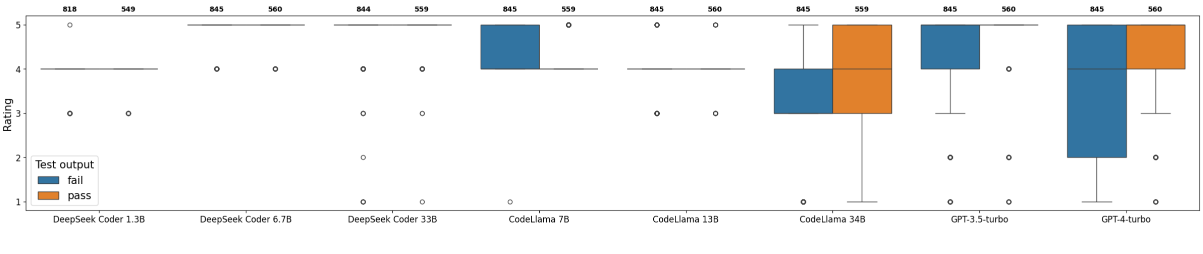

Figure 1: Code Generation **(Java)**: boxplots of judgments provided by the 8 LLMs in the 5-level scenario.

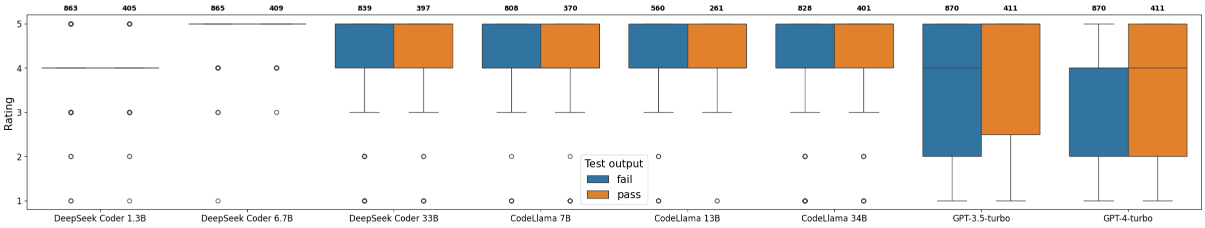

Figure 2: Code Generation **(Python)**: boxplots of judgments provided by the 8 LLMs in the 5-level scenario.